

COMPARAÇÃO DE METODOLOGIAS PARA PREENCHIMENTO DE FALHAS EM SÉRIES HIDROLÓGICAS NA BACIA DO SAPUCAÍ GUAÇU

Camila Billerbeck¹; Silvana Susko Marcellini² & Arisvaldo Vieira Mélo Junior³

RESUMO – A existência de falhas em séries de dados pluviométricos e fluviométricos afetam a qualidade dos resultados de análises hidrológicas e de recursos hídricos. Neste contexto, técnicas de preenchimento de falhas possuem grande importância. Neste trabalho são comparadas metodologias de clássicas de Vetor Regional para preenchimento de dados pluviométricos, e de Regressão Linear para dados fluviométricos, à metodologias baseadas em Redes Neurais Artificiais (RNAs). A metodologia proposta é aplicada em dados mensais de postos da Bacia do Sapucaí-Guaçu. É proposta uma metodologia alternativa para a configuração da RNA, que possibilite o preenchimento simultâneo de falhas nas séries de entrada do modelo. Os modelos de Vetor Regional e RNA obtiveram bons parâmetros de correlação, NSE e RMSE, com relação aos dados observados disponíveis. A RNA obteve melhores resultados para o preenchimento de falhas de dados fluviométricos do que a Regressão Linear. Ambas as metodologias preservaram a sazonalidade temporal das séries satisfatoriamente, tanto para a série de precipitações, quanto para os resultados de vazões.

ABSTRACT– Faults in rainfall and streamflow series affect the quality of the results of hydrological and water resources studies. In this context, gap-filling techniques are of great importance. In this work, Regional Means for gap-filling of rainfall series, and Linear Regression for streamflow series are compared to methodologies based on Artificial Neural Networks (ANNs). The proposed approach is used in the monthly series of the Sapucaí-Guaçu Basin. An alternative methodology for ANN is proposed, which allows the simultaneous gap-filling of faults in the input series of the model. Regional Means and ANN models obtained good correlation parameters, NSE and RMSE, compared to the observed data. RNA obtained better results for gap-filling of streamflow than Linear Regression. Both methodologies preserved the seasonality of the series, both for the series of precipitations and for the results of flows.

Palavras-Chave – Vetor Regional, Regressão Linear, Redes Neurais Artificiais.

1. INTRODUÇÃO E OBJETIVOS

A questão da disponibilidade de séries de dados pluviométricos e fluviométricos é fundamental para análises na área de recursos hídricos. Contudo, comumente encontram-se falhas nos registros cujas causas podem estar relacionadas desde a ausência do observador, falhas nos mecanismos de registro, ou perda das anotações ou transcrições (Oliveira *et al.*, 2010). Portanto, técnicas de preenchimento de falhas possuem grande relevância.

A literatura apresenta algumas metodologias, desde as mais simplificadas até as com maior complexidade, para preenchimento de falhas em séries de precipitação e de vazão em diversas escalas temporais. De maneira geral, as metodologias baseiam-se em métodos estatísticos ou algoritmos que

1) LabSid-PHA-Escola Politécnica (USP): camila@labsid.eng.br / camila.billerbeck@usp.br

2) LabSid-PHA-Escola Politécnica (USP): ssmarcel@usp.br / silvana@labsid.eng.br

3) LabSid-PHA-Escola Politécnica (USP): arisvaldo@usp.br

relacionam parâmetros baseados nas séries observadas, para gerar os dados faltantes (Bier e Ferraz, 2017).

As metodologias disponíveis para preenchimento de falhas em séries de precipitação envolvem ponderação regional, regressão linear, ponderação regional com base em regressões lineares e vetor regional, ou técnicas mais sofisticadas como modelos autorregressivos e de otimização (Oliveira et.al, 2010; Bier e Ferraz, 2017). No caso de preenchimento de falhas em séries de vazão, as metodologias mais utilizadas envolvem transferência de informação de um posto para outro próximo por meio de ponderações, interpolações, métodos estatísticos, regressões lineares, e modelos chuva-vazão (Harvey, *et al.*, 2012; Zhang e Post, 2018). Os métodos de Redes Neurais Artificiais (RNAs) vem ganhando destaque em estudos de recursos hídricos, entre eles, em análises voltadas ao preenchimento de falhas em séries hidrológicas (Teegavarapu e Chandramouli, 2005).

Nesse contexto, este artigo compara metodologias clássicas de preenchimento de falhas para dados mensais de precipitação e de vazão, com metodologias baseadas em RNAs, em postos pluviométricos e fluviométricos na bacia do Sapucaí Guaçu. As metodologias clássicas consideradas nesta análise são vetor regional e regressão linear, para preenchimento de falhas dos postos pluviométricos e fluviométricos, respectivamente. Os resultados obtidos pelos modelos clássicos são comparados aos resultados da RNA. As séries preenchidas resultantes são analisadas em termos de precipitações médias mensais e precipitações totais anuais, e em termos de vazões médias mensais.

2. METODOLOGIA

2.1. Vetor Regional

Formulado por Hiez (1977), este método assume a premissa de que os índices pluviométricos de uma série cronológica podem ser representados por verossimilhança, pelas informações contidas nos dados do conjunto de postos pertencentes a uma região hidrologicamente homogênea. Desta forma, a estimativa da altura precipitada corresponde ao produto mostrado na Equação 1:

$$P_{calc_{ij}} = L_i \cdot C_j \quad (1)$$

Onde $P_{calc_{ij}}$ corresponde à matriz cujos elementos representam a precipitação calculada do mês i no posto j . L_i é o vetor regional, de dimensão i referente ao tamanho da série, e C_j corresponde a um vetor coluna com j elementos, referente ao número de postos pluviométricos da análise. O método consiste em determinar os vetores L_i e C_j de modo que seu produto se aproxime das observações disponíveis nos i postos da região selecionada. Portanto, a função objetivo é minimizar o erro entre as alturas de precipitação observadas e calculadas, conforme a Equação 2:

$$F_{OBJ}(VR) = \min \left\{ \sum_{i=1}^m \sum_{j=1}^n (P_{obs_{mn}} - P_{calc_{mn}}) \right\} \quad (2)$$

Onde $P_{obs_{mn}}$ corresponde à matriz de precipitações observadas.

2.2. Regressão Linear Simples

Essa metodologia utiliza uma correlação entre o posto fluviométrico estudado e um posto fluviométrico vizinho pertencente à uma região hidrologicamente homogênea, cuja relação pode ser descrita linearmente na forma da Equação 3:

$$P_x = aP_y + b \quad (3)$$

Onde P_x é a vazão faltante do posto x; a é o coeficiente angular; P_y é a vazão do posto vizinho y e b é o coeficiente linear da reta. Em casos de mais de dois postos, pode-se utilizar a regressão linear múltipla.

2.3. Redes Neurais Artificiais

Redes Neurais Artificiais (RNAs) são modelos computacionais inspirados no sistema nervoso biológico, com objetivo de mimetizar o funcionamento do cérebro a partir da propagação de dados através de uma rede interconectada. As RNAs possuem características como o compartilhamento de informações, e capacidade de aprendizado. Estes modelos são capazes de reconhecer e mapear padrões entre dados, podendo estendê-los a outro conjunto (Govindaraju e Rao, 2000; Cancelliere *et al.*, 2002; Haykin, 2008).

As unidades fundamentais de uma RNA são os neurônios, cuja ativação ou inibição é regida pela interligação entre seus elementos e seus respectivos pesos. Os neurônios são dispostos em camadas de entrada, de saída e por camadas ocultas, onde ocorre a grande parte dos cálculos (Cancelliere *et al.*, 2002; Haykin, 2008). A Figura 1 mostra um esquema de uma RNA e seus elementos.

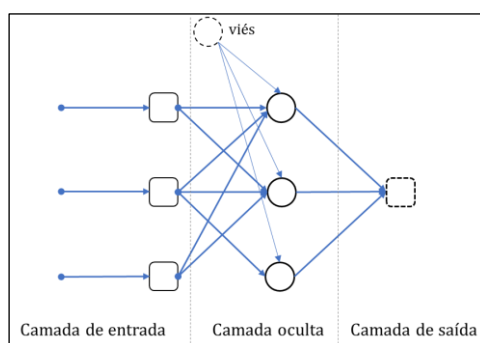


Figura 1 – Esquema de uma RNA e seus elementos.

Para RNAs, recomenda-se a normalização dos dados de entrada, conforme a Equação 4, onde X_i é a variável a ser normalizada; e X_{max} e X_{min} são os valores máximos e mínimos das variáveis:

$$X_{iN} = \frac{(X_i - X_{min})}{(X_{max} - X_{min})} \quad (4)$$

Estão disponíveis inúmeros algoritmos para cálculos de RNAs. Em particular, o pacote *neuralnet* (Günther e Fritsch, 2010), permite flexibilidade em que o usuário define parâmetros como o número de neurônios na camada oculta, função de ativação, número máximo de iterações e função de erro. Neste trabalho o pacote *neuralnet* foi utilizado por meio da interface *RStudio*.

Quanto à arquitetura das RNAs em aplicações para preenchimento de falhas é usual a adoção de postos vizinhos para preencher falhas em um determinado posto específico, ou seja, uma RNA com múltiplas camadas de entrada e uma camada de saída (Ledra, *et al.*, 2017; Correia, *et al.*, 2016). A desvantagem é a quantidade de redes neurais a serem calibradas, caso haja mais de um posto com falhas. Este procedimento também apresenta variações quanto à seleção da série temporal para calibração da RNA, pode-se calibrar doze RNAs representando os meses das séries disponíveis, ou uma RNA para os anos hidrológicos disponíveis.

A abordagem deste trabalho, propõe uma RNA que tem como entrada os dados dos postos selecionados contendo ou não falhas a serem preenchidas, e que resulte em séries geradas para todos os postos de entrada. O objetivo desta abordagem é obter uma única rede neural para o preenchimento de falhas simultâneas nos postos. Para isso a metodologia foi organizada em três etapas: (i) seleção do período de dados para calibração, (ii) calibração e validação da RNA, e (iii) geração de séries preenchidas. A Figura 2 mostra um esquema da metodologia descrita.

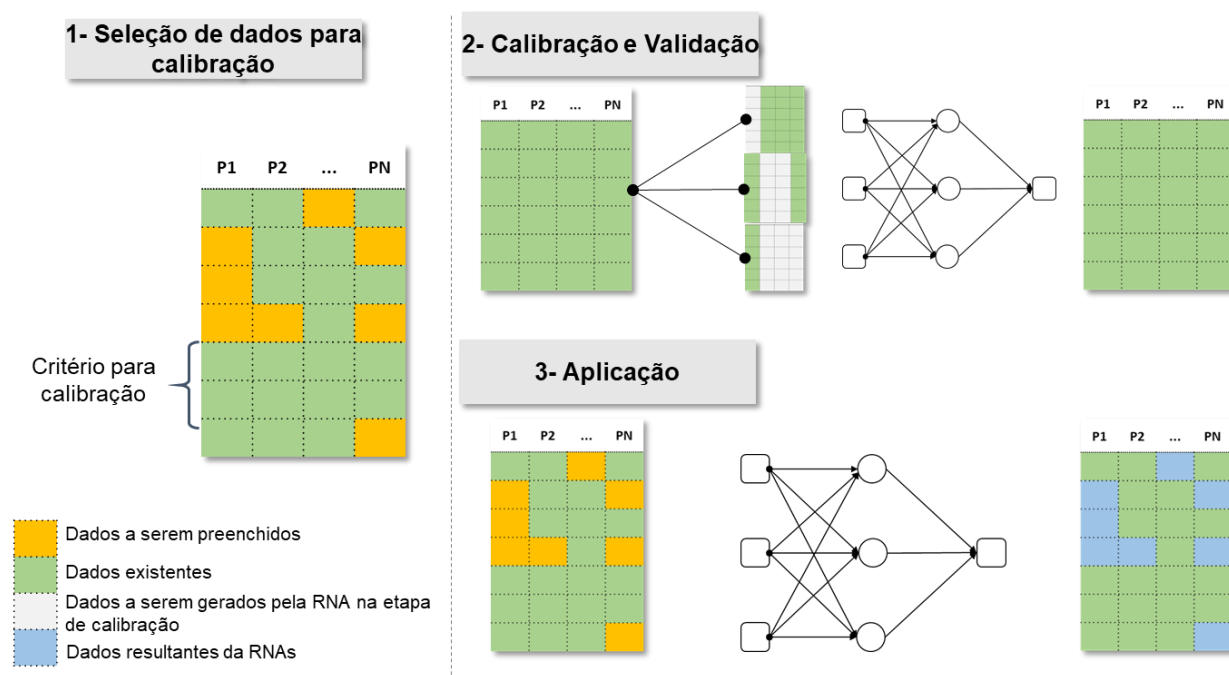


Figura 2 – Esquema da metodologia proposta para preenchimento de falhas de postos fluviométricos e pluviométricos com RNA, onde: P1,P2.. Pn – representam os postos selecionados em regiões hidrológicamente homogêneas.

Na etapa (i) seleciona-se períodos com dados concomitantes em todos os postos disponíveis para a calibração da RNA. A etapa (ii) envolve o preparo dos dados de entrada para o treinamento da RNA, e a calibração do modelo. Nesta etapa é necessário preparar o arquivo de entrada que consiste em uma série resultante da combinação entre o número de postos tomados como referência para gerar os dados faltantes. Quanto maior o número de postos disponíveis na análise, maior a série proveniente da combinação conforme mostra a Equação 5.

$$\sum_{p=1}^{n-1} C_{n,p} = \frac{n!}{p!(n-p)!} \quad (5)$$

Por exemplo, supondo três postos ($n=3$), as combinações consideradas para entrada no modelo são: três combinações de um posto, para gerar dados para um par de postos ($p=2$), mais três combinações de dois postos, para gerar dados para um posto ($p=1$).

A arquitetura de RNA, que gerou melhor convergência consiste em uma camada de entrada, três camadas ocultas e uma camada de saída. Os parâmetros de entrada são os valores observados, um parâmetro referente ao mês, e um indicador qualitativo adicional que identifica se o valor de entrada correspondente deve ser preenchido ou tomado como referência. Este indicador foi assumido associado a valores binários na etapa de preparação do arquivo de entrada. A Figura 3 apresenta um esquema da estrutura das RNAs formuladas.

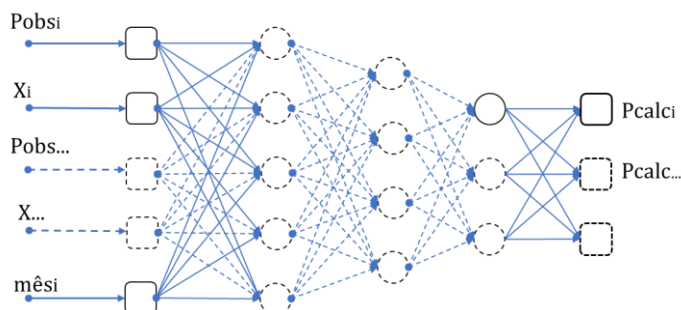


Figura 3 – Esquema da RNA para preenchimento de falhas, onde: *Pobs* – valor observado no posto no mês *i*, *X* – indicador qualitativo, *mês* – indicador de mês da série, *Pcalc* – valor obtido pela RNA.

3. ESTUDO DE CASO: BACIA DO SAPUCAÍ GUAÇU

A Tabela 1 apresenta os postos pluviométricos e fluviométricos da bacia do Sapucaí-Guaçu selecionados para o presente estudo, e a Figura 4 mostra sua localização.

Tabela 1 – Localização e Características dos postos selecionados na bacia do Sapucaí-Guaçu.

Código ANA	Tipo	Nome	Coordenadas (Lat, Long)	Período	Período considerado para preenchimento de falhas
2245010	Pluviométrico	Fazenda da Guarda	-22,68;-45,48	1941-2017	1937-2017
2245018	Pluviométrico	Vila Capivari	-22,72;-45,57	1937-2017	1937-2017
2245022	Pluviométrico	Usina do Fojo	-22,71;-45,53	1937-1997	1937-2017
2245163	Pluviométrico	Marmelos	-22,63;-45,50	1971-1998	1937-2017
2245167	Pluviométrico	Campos do Jordão	-22,70;-45,48	1972-2002	1937-2017
61250000	Fluviométrico	Fazenda da Guarda	-22,69;-45,48	1934-2014	1934-2014
61250100	Fluviométrico	Parque Estadual	-22,69;-45,48	1980-2018	1934-2014

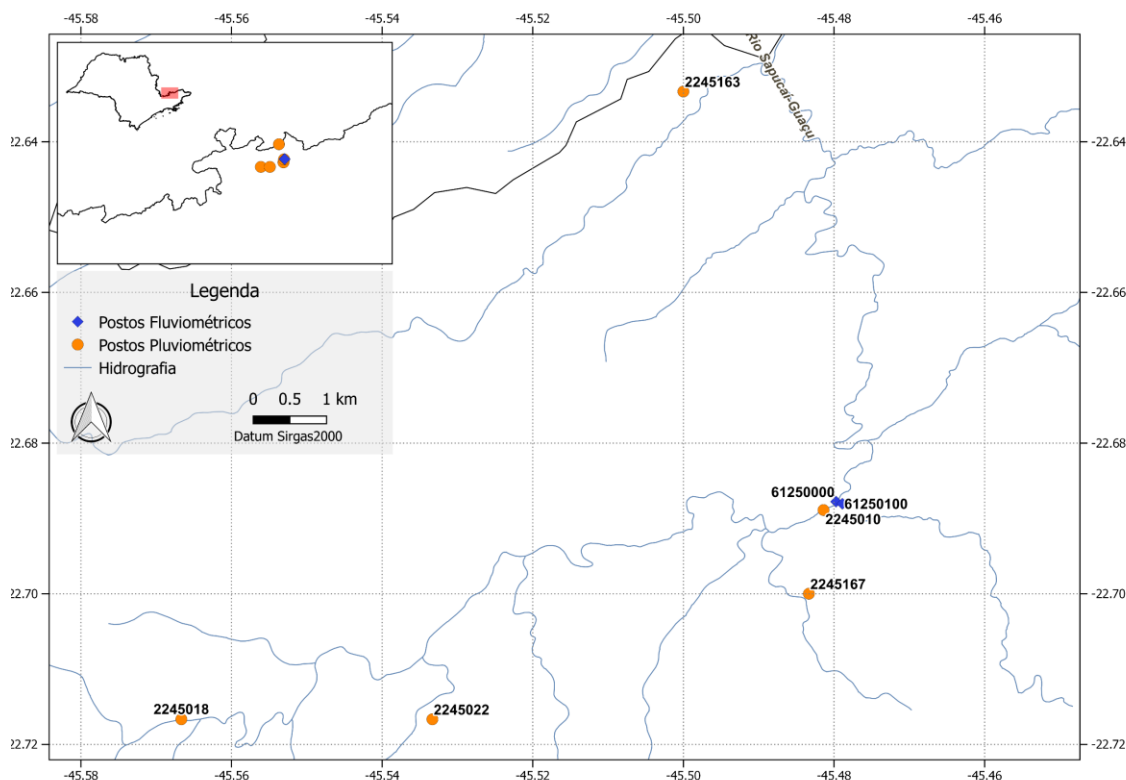


Figura 4 – Postos Pluviométricos e Fluviométricos da bacia do Sapucaí-Guaçu selecionados.

Para avaliar a calibração dos modelos de rede neural, vetor regional e regressão linear, foram analisados os coeficientes de correlação (r), a raiz do erro quadrático médio (RMSE), o coeficiente de eficiência de Nash-Sutcliffe (NSE), mostrados respectivamente nas equações 6, 7 e 8. A RMSE fornece uma ordem de grandeza dos erros de predição do modelo, sendo desejáveis valores mais próximos de zero, e o coeficiente NSE varia de $-\infty$ a 1, sendo quanto mais perto de 1, melhor a eficiência do modelo.

$$r = \frac{\sum_{i=1}^N (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{(\sum_{i=1}^N (X_i - \bar{X})^2) \cdot (\sum_{i=1}^N (Y_i - \bar{Y})^2)}} \quad (6)$$

$$RMSE = \left[\frac{1}{N} \cdot \sum_{i=1}^N (X_i - Y_i)^2 \right]^{1/2} \quad (7)$$

$$NSE = 1 - \frac{\sum_{i=1}^N (X_i - Y_i)^2}{\sum_{i=1}^N (X_i - \bar{X})^2} \quad (8)$$

Onde: \bar{X} e \bar{Y} são as médias dos respectivos conjuntos de dados; e N o número de elementos no conjunto.

4. RESULTADOS E DISCUSSÃO

A seguir são apresentados os resultados obtidos para as metodologias descritas.

4.1. Preenchimento de falhas de Precipitações

Os cinco postos pluviométricos selecionados na bacia do Sapucaí-Guaçu abrangem períodos distintos de disponibilidade de observações, para o preenchimento de falhas foi considerado o período mais longo, entre 1937 e 2017.

A metodologia do vetor regional foi aplicada para todos os dados disponíveis neste período, diferentemente da RNA que foi calibrada com base no período de observações concomitantes, entre os anos de 1972 a 1998, aproximadamente. A RNA formulada teve 11 parâmetros de entrada, referentes às observações de cada posto, o indicador qualitativo e o indicador de mês, três camadas ocultas contendo 11, 8 e 5 neurônios respectivamente, e com 5 parâmetros de saída representando os valores de precipitação calculados.

A princípio, foi formulada uma única RNA para todos os dados de calibração no período, o que resultou em precipitações elevadas entre os meses mais secos, o que mostrou a necessidade de separar os períodos hidrológicos. Sendo assim, foram calibradas duas RNAs, uma para os meses chuvosos, entre outubro e março, e outra para os meses mais secos entre abril e setembro.

A Tabela 3 mostra os mesmos parâmetros obtidos com relação às séries de precipitação mensal disponíveis sem o preenchimento de falhas, para as metodologias do Vetor Regional e RNA.

Tabela 3 – Correlação, RMSE e NSE obtidos para a RNA e Vetor Regional quanto à precipitação mensal observada.

Parâmetro	Posto 1 2245010		Posto 2 2245018		Posto 3 2245022		Posto 4 2245163		Posto 5 2245167	
	Vetor Regional	Rede Neural	Vetor Regional	Rede Neural	Vetor Regional	Rede Neural	Vetor Regional	Rede Neural	Vetor Regional	Rede Neural
correlação	0,94	0,96	0,94	0,92	0,94	0,98	0,94	0,97	0,98	0,98
RMSE	41,7	36,7	43,9	51,2	41,9	26,6	42,2	29,5	26,3	25,5
NSE	0,89	0,91	0,88	0,84	0,88	0,95	0,88	0,98	0,95	0,98

Os parâmetros de correlação e NSE próximos de 1 indicam que ambas as metodologias apresentaram bons resultados para estimar os valores observados. Os valores de RMSE indicam a ordem de grandeza aproximada do erro entre os valores calculados e observados, em geral o método do Vetor Regional obteve variabilidade de RMSEs entre 41,7 e 43,9 mm para os postos 1 a 4, e 26,3 mm para o posto 5. Enquanto a RNA apresentou RMSEs entre 26,6 e 51,2 mm, e igual a 25,5 mm, respectivamente. O Posto 5 apresentou os menores RMSEs, e coeficientes de correlação e NSE maiores ou iguais dos verificados nos demais postos, para as duas metodologias.

O Posto 2, que possui a maior disponibilidade de dados observados, obteve melhores resultados para os coeficientes com o modelo do Vetor Regional do que pela RNA. Este resultado pode indicar que é provável que o principal posto para a transferência de resultados para os demais, tenha sido o Posto 2, por este modelo. O mesmo não ocorre com a RNA, pois as séries utilizadas para calibração foram escolhidas de acordo com a disponibilidade de dados em períodos concomitantes. O melhor ajuste para a RNA em termos dos parâmetros apresentados foi o Posto 5, que possui a série com maiores falhas.

A Figura 5 apresenta um exemplo da distribuição sazonal das precipitações médias mensais obtidas para o posto Marmelos (2245163) preenchido pelas metodologias do Vetor Regional e da RNA. Nos meses chuvosos, entre Outubro e Março, os valores de precipitação média obtidas pela RNA superam os resultados médios obtidos pelo Vetor Regional, no entanto, a magnitude dos resultados é próxima e a sazonalidade é refletida em ambos os modelos.

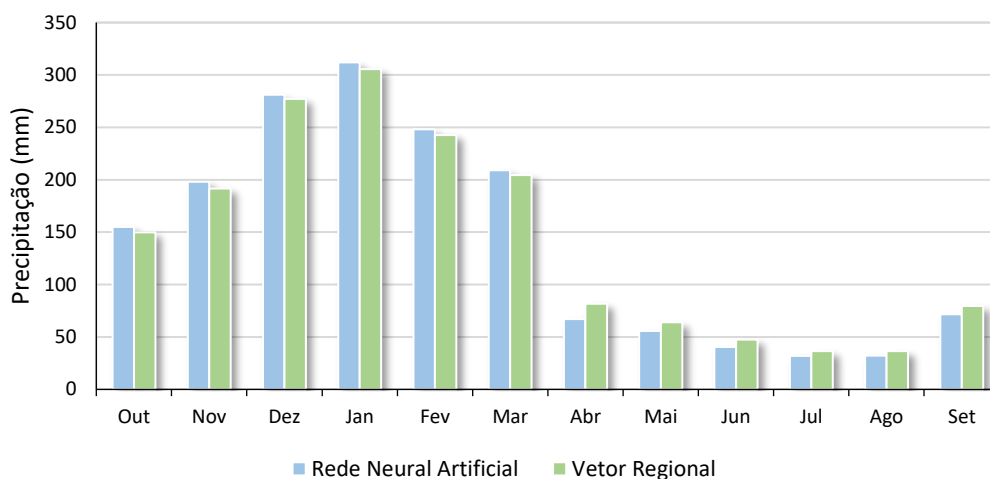


Figura 5 – Distribuição sazonal de precipitações médias mensais para o posto Marmelos (2245163).

A Figura 6 apresenta as precipitações totais anuais da série entre 1937 e 2014, com dados preenchidos com o Vetor Regional e RNA, no posto Marmelos (2245163). Pode-se observar que as séries possuem comportamentos e sazonalidade próximos para todo o período. Este comportamento também é observado nos demais postos preenchidos, para ambos os modelos.

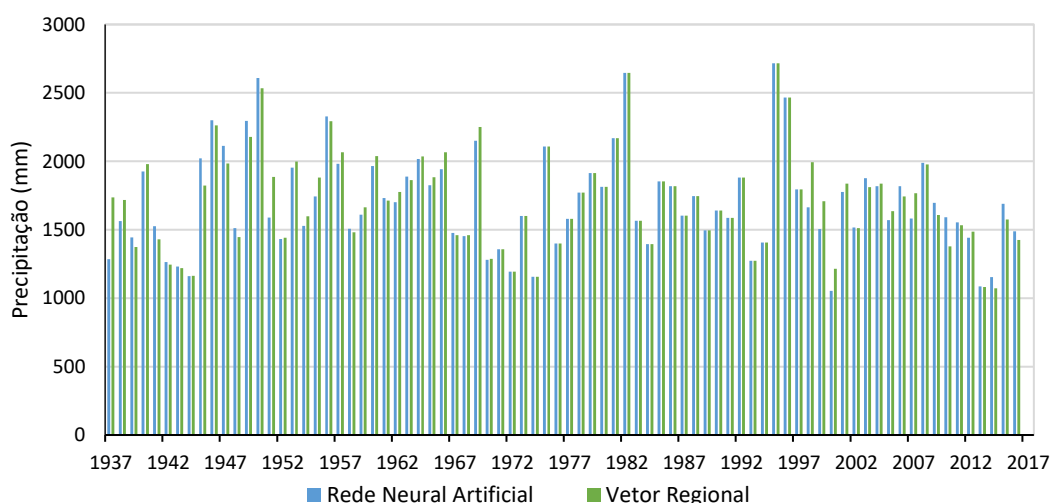


Figura 6 – Precipitações totais anuais, com valores preenchidos com o vetor Regional e RNA, para o posto Marmelos (2245163).

4.2. Preenchimento de falhas de Vazões

No caso dos postos fluviométricos, o posto Fazenda da Guarda (61250000) com dados disponíveis entre 1935 e 2014 apresenta poucas falhas, já o posto Parque Estadual (61250100)

apresenta maiores falhas em observações e menor período, entre 1980 e 2014, portanto procedeu-se também a extensão da mesma entre os anos de 1935 e 1980.

A RNA formulada teve 5 parâmetros de entrada, referentes às observações de cada posto, o indicador qualitativo e o indicador de mês, três camadas ocultas contendo 5, 3 e 2 neurônios respectivamente, e com 2 parâmetros de saída representando os valores de vazão obtidos. Ambas as metodologias foram calibradas com dados de observações concomitantes.

A Tabela 4 mostra os parâmetros de correlação, RMSE e NSE obtidos na calibração dos métodos de regressão linear e RNA para os postos fluviométricos.

Tabela 4 – Parâmetros de correlação, RMSE e NSE obtidos na calibração das vazões médias mensais dos métodos de regressão linear e RNA.

Parâmetro	61250000		61250100	
	Regressão linear	RNA	Regressão linear	RNA
correlação	0,92	0,94	0,92	0,94
RMSE	0,64	0,22	0,56	0,29
NSE	0,85	0,88	0,85	0,89

O modelo da RNA obteve melhor calibração com os dados observados se comparada à regressão linear, com maiores parâmetros de correlação e NSE, e menor RMSE.

A Figura 7 mostra a distribuição sazonal de vazões médias mensais do posto Parque Estadual (61250100), que teve o maior número de falhas preenchidas, obtidas por meio de regressão linear com o posto Fazenda da Guarda (61250000) e pela RNA. Percebe-se que as vazões médias mensais resultantes da regressão linear tendem a ser maiores do que as obtidas pela RNA em todos os meses. Em ambos os modelos, o comportamento sazonal das vazões médias foi representado.

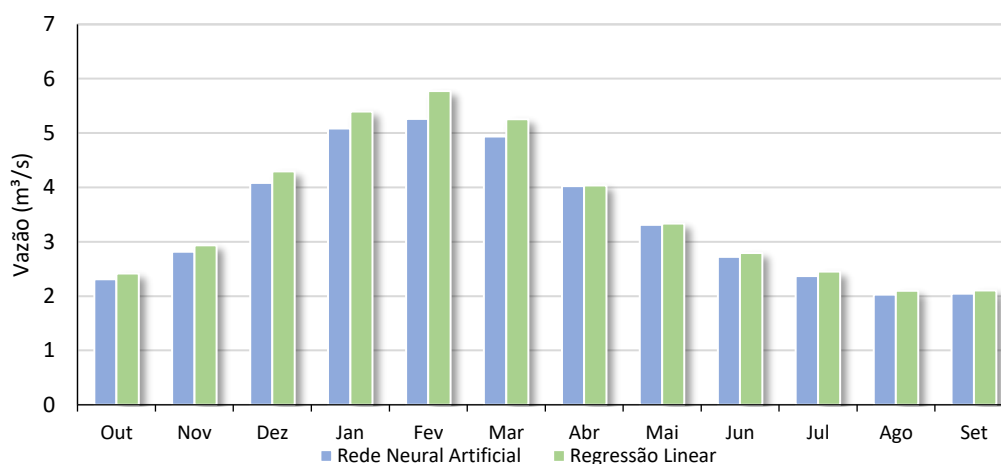


Figura 7 – Vazões médias mensais para o posto Parque Estadual (61250100).

5. CONCLUSÕES

Compararam-se metodologias clássicas de preenchimento de falhas em dados pluviométricos (Vetor Regional) e fluviométricos mensais (regressão linear) com a metodologia baseada em RNAs

na Bacia do Sapucaí-Guaçu. A proposta de realizar combinações simples para os dados de calibração da RNA mostrou-se como uma alternativa viável, possibilitando que todos os postos disponíveis sejam fornecidos como dados de entrada e de saída.

Para o preenchimento de falhas das séries de precipitação foi necessária a formulação de duas RNAs para representar períodos chuvosos, entre outubro e março, e secos, entre abril e setembro. Os modelos de Vetor Regional e RNA obtiveram bons parâmetros de correlação, NSE e RMSE, com relação aos dados observados disponíveis, para os cinco postos. O Posto 2, que possui o maior período de dados observados, obteve melhores resultados com o modelo do Vetor Regional do que pela RNA. O melhor resultado para a RNA em termos dos parâmetros foi no Posto 5, que possui uma série com maiores falhas. Ambos os modelos preservam a sazonalidade das precipitações mensais médias e totais anuais.

Com relação ao preenchimento de falhas nos postos fluviométricos, a RNA obteve melhor calibração com o conjunto de dados de referência, e menores vazões médias mensais do que a regressão linear. Ambos os métodos se mostraram capazes de preservar o comportamento sazonal das vazões médias mensais.

REFERÊNCIAS

- BIER, A.A.; FERRAZ, S.E.T. (2017). “*Comparação de Metodologias de Preenchimento de Falhas em Dados Meteorológicos para Estações no Sul do Brasil*”. Revista Brasileira de Meteorologia. (32-2), pp. 215-226.
- CANCELLIERE, A. *et al.* (2002). “*A Neural Networks Approach for Deriving Irrigation Reservoir Operating Rules*”. Water Resources Management (16-1), pp. 71-88.
- CORREIA, T.P. *et al.* (2016). “*Aplicação de Redes Neurais Artificiais no Preenchimento de Falhas de Precipitação Mensal na Região Serrana do Espírito Santo*”. Geociências, (35-4), pp. 560-567. Geografia Física Aplicada, Campinas, Jul. 2017, 1, pp. 1747-1755.
- GOVINDARAJU, R.S.; RAO, A.R. (2000). “*Introduction*” in *Artificial Neural Networks in Hydrology*. Org por Govindaraju, R.S.; Rao, A.R. Springer Netherlands.
- HARVEY, C.L.; DIXON, H.; HANNAFORD, J. (2012). “*An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK*”. Hydrology Research. (43-5), pp. 618-636.
- HAYKIN, S.O. (2008) “*Neural Networks and learning machines*. 3 ed. Pearson New Jersey.
- HIEZ, G. (1977). “*L’homogénéité des données pluviométriques*”. Cahiers ORSTOM. Série Hydrologie. (14-2), pp.129-172.
- LEDRA, L.; BARROS, V.G; HENNING, E. (2017). “*Preenchimento de Falhas de Dados Diários de Precipitação Utilizando Redes Neurais Artificiais e Método das Correlações*” in Anais do XXII Simpósio Brasileiro de Recursos Hídricos, Florianópolis, Dez. 2017.
- OLIVEIRA, L.F.C.; FIOREZE, A.P.; MEDEIROS, A.M.M; SILVA, M.A.S. (2010). “*Comparação de metodologias de preenchimento de falhas de séries históricas de precipitação pluvial anual*”. Revista Brasileira de Engenharia Agrícola e Ambiental, (14-11), pp. 1186–1192.
- TEEGAVARAPU, R.S.V.; CHANDRAMOULI, V. (2005). “*Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records*”. Journal of Hydrology. (312), pp. 191-206.
- ZHANG, Y.; POST, D. (2018). “*How good are hydrological models for gap-filling streamflow data?*”. Hydrology and Earth Systems. (2-8), pp. 4953-4604.